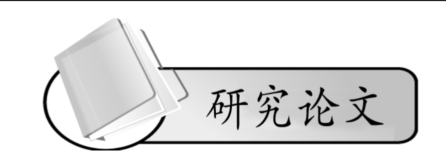


基于 WMD 语义相似度的 TextRank 改进算法识别论文核心主题句研究

王子璇 [1, 2]　乐小虬 [1]　何远标 [1]

[1](中国科学院文献情报中心　北京　100190)

[2](中国科学院大学　北京　100049)

摘要:【目的】自动甄别科技论文中描述研究主题的关键语句。【方法】以论文小节为单位组织句子集，通过训练领域词向量计算句子间 WMD 距离得到相应语义相似度，优化 TextRank 算法迭代过程，利用外部特征对所得权值进行调整，按句子权值降序选取关键主题句。【结果】以气候变化领域科技论文作为实验数据，以人工标注的结果为基准对本文的算法和传统的TextRank算法进行对比实验，初步结果表明该方法的识别效果(F值)比传统 TextRank 算法提升约 5%。【局限】句子特征提取有待提高，词向量训练及方法中的相关参数需要做进一步优化。【结论】基于领域词向量，融合 WMD 语义相似度的 TextRank 改进算法，能够较好地甄别科技论文小节内部中心句，辅以外部特征的权值调整后可以较好地识别出一篇论文的核心主题句。

关键词: WMD　TextRank　语义相似　主题句识别　外部特征

分类号: TP393

1　引　言

科技论文中作者常聚焦于一个主要研究问题，在文献分析中可用研究主题来表示，主题句是论文中用于论证研究主题的句子，分布于文中主要段落中。主题句识别作为文本分析的基础技术之一，其在信息检索、自动文摘及知识发现等自然语言处理应用中发挥着重要作用。识别领域科技论文中的核心主题句，就是要从全文中将描述和揭示研究主题的关键语句进行鉴别和抽取。它是科技论文内容提炼的关键技术环节，能帮助研究者快速发现论文中相对有价值的内容，提高科研效率。

文本主题句识别的一般过程为：识别文本中的候选主题句；合理评估这些候选主题句表达文本核心内容及其主题的重要程度，从中挑选合适的句子作为主题句[1]。而评估句子重要性的方法主要是通过度量句子自身所带特征(位置、主题词、长度等)以及句子之间的相互关系进行评估。前者主要利用自身统计特征构建模型进行权值打分或监督学习，而后者则将句子及其关系转化为图模型进行识别，以 TextRank[2]为代表。

传统 TextRank 中使用特征词向量表示句子，再利用距离相似度计算方法(如欧氏距离、余弦相似度等)计算句子间相似度，但在句子表示上存在维数灾难及同近义词的问题。为了解决以上问题，本文将以基于词向量(Word Embedding)语义相似的 WMD(Word Mover's Distance)[3]表示句子间的距离，对 TextRank 算法进行改进，并利用论文内容结构对所得结果进行优化，更新权重并排序，最终得到科技论文的核心主题句。

2　主题句识别相关研究

主题句识别作为多项自然语言处理应用的基础

通讯作者：乐小虬，ORCID：0000-0002-7114-5544，E-mail：lexq@mail.las.ac.cn。

性工作，国内外学者对此提出了多种方法，因时间发展和技术手段不同，主要有以下三种:

(1) 基于统计特征的方法。通过将原文本转变为句子的线性序列，把句子转变为词的线性序列，以一定的特征指标对词语句子赋予相应的权重，最终选择综合权值较高的句子作为输出，得到主题句[4]。Luhn[5]指出的词频、Baxendale[6]提及的句子位置，以及刘挺等[7]总结的标题、位置、句法结构等信息均可作为衡量句子重要性的指标。Edmundson[8]选择其中的几种变量，构造了一个简单的多元线性函数：$Weight(x)=a_1C+a_2K+a_3T+a_4L$，其中 C、K、T、L 分别为 4 种特征变量，其他为调节参数，用多种特征描述句子的权重值。实践表明这种表示方法并不理想，其线性相加的过程缺乏理论基础。统计特征的方法具有过程简单、识别速度快的特点，但特征选择及加权方法很大程度影响识别结果，效果并不是很稳定。

(2) 基于机器学习二分类的方法。将文本主题句的识别转换句子层面是否是主题句的二分类问题进行判别，其主要包括特征选择、算法选择、模型训练、主题句判别 4 个步骤，可用于主题句识别的机器学习算法有朴素贝叶斯、条件随机场、支持向量机等多种模型。Kupiec 等[9]以句长、固定短语、段落、主题词和大写字母词语等特征，首次运用 NB 分类器对文本进行主题句识别。Conroy 等[10]将 HMM 运用于主题句识别，利用观测序列寻找最可能的隐含状态序列，将句子位置、词频及词的概率 3 个文档特征构成观测状态转移概率矩阵，构建预测模型判别。机器学习二分类的方法效果虽然较好，但需要提前准备较多训练数据，且依赖于特征独立性假设，适用性和可操作性不强。

(3) 基于图排序的方法。将文本段分解为若干个句子单元，每一个单元对应图结构中的一个顶点，各个单元之间的相似性关系作为边，通过图排序的算法得出各顶点的得分，并在此基础上选择较高得分的句子作为主题句。不同的图排序方法主要因边权值的计算方式以及图排序算法的选择不同。常见的边权值计算方式包括词共现、句子相似度等；图排序算法包括矩阵权值相加、PageRank 等方法。Mihalcea 等[2]首先提出 TextRank 算法，将图排序应用到主题句识别中，并在文献[11]中做了优化。余珊珊等[12]过滤非重要词，归并同近义词，采用向量空间模型表示句子，计算余弦相似度作为边的权值，并加入人工特征对 TextRank 算法计算结果进行优化。而耿焕同等[13]、何维等[14]利用词共现形成的主题信息以及不同主题间的连接特征识别主题句。图排序方法不需要外部知识以及训练样本就可取得结果，但结果质量受边权值计算方式的影响，也不是很稳定。

对于科技论文这种内容结构化较强的文本数据，句子间存在隐含的语义联系。因此，本文在图排序算法进行主题句识别的基础上，对句子间相似度计算方法进行改进，并考虑文本结构化特征对结果进行优化，提升识别结果。

3 识别方法

科技论文是作者对科研过程相关研究的总结，Saïd 等[15]指出作者的写作思路在一定程度上影响了文章的内容结构，这种结构体现在文章的逻辑元素(如标题、段落、片段等)之间的关系。同时利用文本段落句子的内部联系以及论文整体的外部结构对识别文本段落中的核心成分将会起到重要作用。

本文提出的论文核心主题句识别方法主要包括词向量表示与训练、句子间相似度计算、TextRank 算法迭代计算、利用外部结构特征优化结果 4 个步骤，如图 1 所示。

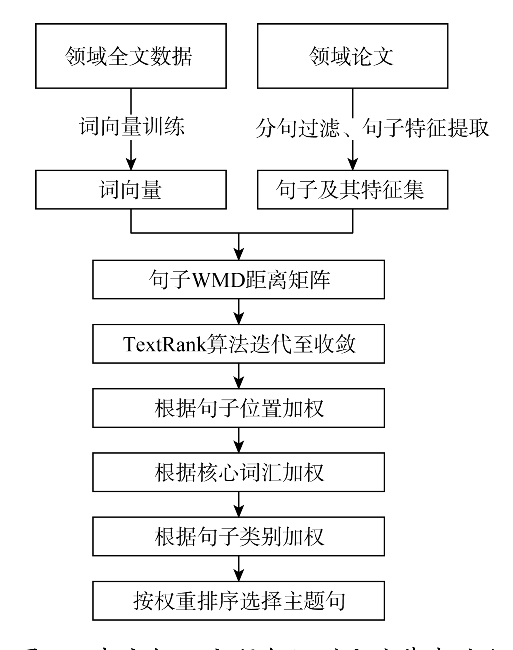


图 1 本文核心主题句识别方法基本流程

首先，利用 Word2Vec 模型对领域科技论文全文进行训练，得到领域词向量；其次，对各个论文文本段落进行分句，去除无意义的短句，利用训练好的词向量表示句子间的 WMD 距离，并转换为句子相似度；再者，针对每个文本段落构建无向权重图，边权重用句子间相似度表示，用 TextRank 算法进行迭代计算，得到各句子的权重值；最后，利用句子位置、大纲结构等特征信息对权重值进行调整并排序，最终按比例识别出论文的核心主题句。

3.1 基于 WMD 语义相似度的 TextRank 改进算法

(1) 词向量表示与训练

词是承载语义的最基本的单元。传统的独热表示(One-hot Representation)把每个词孤立，并用 0 和 1 表示，整个向量并不包含语义信息，并且存在维数灾难问题。Harris[16]提出的分布假说(Distributional Hypothesis)表明词的语义由其上下文决定。Bengio 等[17]提出神经网络语言模型(Neural Network Language Model, NNLM)，通过神经网络语言模型对目标词以及更复杂的上下文之间关系进行建模，在学习语言模型的同时，也得到了维数较低的副产品——Word Embedding，俗称词向量。这种传统 NNLM 模型计算复杂度较高，对于较大数据集运行效率低。Mikolov 等[18]在此基础上移除了隐藏层，提出了 CBOW (Continuous Bag-of-Words)和 Skip-gram 模型。CBOW 模型是用上下文的词预测该词，Skip-gram 则相反，以当前词的词向量为输入，输出层是该词周围单词的词向量。而训练过程则有两种优化方法：Hierarchical Softmax，通过将原 Softmax 方法转换带有霍夫曼树的层级 Softmax，利用霍夫曼树的特性将预测时间缩短 $\log n$ 倍；Negative Sampling，利用更简单的随机带权负采样方法可以大幅度提高性能，加快计算速度。这两种模型与方法可以任意搭配使用，并且在 Google 于 2013 年发布 Word2Vec[19]开源工具包里有完整实现。

(2) 句子相似度计算

传统的句子相似度计算方法包括：基于特征词的方法，利用 TF-IDF、卡方值、互信息等选择特征词，构建向量进行计算[20]；基于句法分析的方法，对句子进行句法分析，计算句子之间的句法结构及内容的相似程度，目前以简单句法结构匹配为主[21]；基于语义分析的方法，通过语义词典或本体对句子中词语进行消歧，并在计算过程中兼顾词语在语义层面的相似性[22]。前两种方法存在向量维数灾难及同近义词的问题，而第三种方法依赖外部知识，外部知识的好坏及未收录的词直接影响了计算结果，不具有可扩展性。

词向量较好地解决了上述问题，训练过程简单，且得到的词与词之间存在潜在的语义关系，如 Mikolov 等[23]发现两个词向量之间存在着加减关系，$c(king)-c(queen)\approx c(man)-c(woman)$。将句子中每个词的词向量直接相加并做归一化得到句向量，计算句向量间余弦值有不错的效果。Kusner 等[3]在此基础上提出了 Word Mover's Distance(WMD)，词–词相似度用欧氏距离表示，句子–句子相似度转化为运输最优化问题，将两个句子相似看成两个概率分布的变换，其句子间距离由变换代价表示，并应用于 KNN 文本分类取得良好效果。同时，文章证明了对词向量求平均值算欧式距离是 WMD 的下界。

句子间 WMD 距离计算，首先将两个句子 s,s' 转变为词袋，并去除停用词，共计 n 个词，对剩下词的词频做归一化处理，构建词频向量，记为 $d,d'\in R^n$，计算词袋中每两个词的欧式距离作为词转换的运输成本，构建转移矩阵 $T\in R^{n\times n}$，将其中一个句子所有的词转变为另一个句子的所有词所耗费的成本记为 $\sum_{i,j}T_{i,j}c(i,j)$，计算距离转变为求解运输成本最小值问题，如公式(1)所示。

$$\min_{T\geqslant 0}\sum_{i,j=1}^{n}T_{i,j}c(i,j)$$
$$s.t.:\ \sum_{j=1}^{n}T_{ij}=d_i,\forall i\in\{1\cdots n\} \quad (1)$$
$$\sum_{i=1}^{n}T_{ij}=d_j{}',\forall j\in\{1\cdots n\}$$

采用 EMD 算法[24]解决，变换过程如图 2 所示[3]。

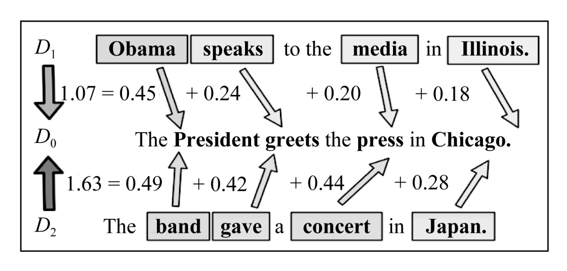


图 2　句子 WMD 距离计算过程

句子间相似度采用类似欧氏距离转相似度的方

法，最终得到句子间的相似度度量，如公式(2)所示。

$$sim(s,s')=\frac{1}{1+wmd(s,s')} \quad (2)$$

(3) TextRank 算法迭代计算

一篇文档通常由多个段落组成，连续的段落在内容上是相近的，形成语义内聚的节，对应一个子主题，并按照大纲层级统一在上层主题下。同一节下内容相近段落中的句子构成独立的句群，以句群为单位构建网络图进行主题句识别，可以很好地识别出能相对代表整个小节内容的主题句，保证识别效果。

TextRank方法借鉴PageRank算法的思想，将句子间的相似关系看成是一种支持或推荐的关系，将句子作为节点，利用句子间相似度表示句子间的关系作为边，构建图模型，通过迭代计算优化各句子的权重值，再选择权重较高的句子作为主题句。

假设文本段落由集合 V 中 n 个句子 $V_i(1\leqslant i\leqslant n)$ 组成，以 V_i 为节点并以节点间相似关系为边构建 TextRank 网络图 G。通过前面句子相似度计算可得到 $n\times n$ 的句子相似度矩阵 $S_{n\times n}$，如公式(3)所示。

$$S_{n\times n}=\begin{pmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nn} \end{pmatrix} \quad (3)$$

根据给定的 G 及 $S_{n\times n}$ 迭代计算各个节点的权重，权重计算如公式(4)所示。

$$WS(V_i)=\frac{1-d}{n}+d\times\sum_{V_j\in In(V_i)}\frac{s_{ij}}{\sum_{V_k\in Out(V_j)}s_{jk}}WS(V_j) \quad (4)$$

其中，$WS(V_i)$ 是节点 V_i 的权重值，$In(V_i)$ 代表指向 V_i 的节点集合，$Out(V_i)$ 代表 V_i 所指向的节点集合，d 是阻尼系数，一般设置为 0.85。各个节点的初始权重一般设为 $1/n$，当两次迭代后权重变化差别非常小并接近于零时停止迭代，最终得到各句子的权值，最后根据权值排序选择一定比例的权值较大句子作为主题句。

3.2 外部特征选取及权值计算

经过迭代计算得到的句子权重会收敛趋于稳定，这个稳定值由其他句子决定，而与初始值无关，故在算法迭代前调整各句子的权值是无意义的。在 TextRank 算法完成后，通过加入外部特征对所得句子权重序列进行调整，具体特征调整方法如下。

(1) 句子所在位置

Baxendale[6]通过统计发现段落的主题句为段落首句的概率为 85%，为段落末句的概率为 7%。同理对小节中段落而言，首尾段落则更有可能对小节内容进行引导或总结，揭示主题内容。故可根据段落在小节中的位置以及句子在段落中的位置对句子权重进行加权，首段和尾段中句子权重提升更大，段落中首句和尾句的权重提升也更大，两种加权方式可采用同样的函数，具体加权函数如公式(5)所示。

$$\begin{cases} 1+(1-x)\times e_1 & 0\leqslant x\leqslant 0.3 \\ 1 & 0.3<x<0.7 \\ 1+x\times e_2 & 0.7\leqslant x<1 \end{cases} \quad (5)$$

其中，e_1 和 e_2 为可调整的阈值，这里可取 $e_1=e_2=0.1$，x 表示句子所在段落位置或段落所在小节的百分比，由前至后分别为 0 至 1，权重提升比分别记为 p_1 及 p_2，最终权值调整为:

$$WS(V_i):=WS(V_i)\times p_1\times p_2 \quad (6)$$

(2) 核心术语

TextRank 算法从句群内部提炼出了相对能表达小节自身内容的句子，但不一定与文章所表达的主题相关。因此，可考虑利用文章标题、关键词、大纲等相对于句群的外部特征来进行优化。利用何远标等[25]提出的大纲术语抽取方法，对标题、小节所在的大纲层级结构进行术语识别并与关键词合并，得到核心术语集合。越能体现核心术语的句子，则越有可能是核心主题句。对于核心术语在句子中的体现，目前的解决方案仅为包含关系，即 $WS(V_i):=WS(V_i)\times(1+p_3\times n)$，$p_3$ 为加权权重，这里取 0.1，n 为包含的核心词汇个数。

(3) 句子类别

文献[26]指出大纲中除了一些具体的术语，也包含大量具有广泛意义的论文术语，如“method”、“conclusion”等，是针对某研究点的分面描述，可作为主题描述框架的标志。对于论文全文内容，具有一定类别的句子则具有更大的价值，能更好地阐述主题相关的分面描述。同时这种带有一定类别句子的识别，正好是结构化摘要的体现。故可考虑对带有一定类别的句子进行加权，而对于含有论文术语的大纲及文本段，对应论文术语类别的句子则是文本段内容的集中体现，可加大该部分句子的加权比重，具体加权函数为 $WS(V_i):=WS(V_i)\times[1+p_4\times(n+5\times b)]$，其中 p_4 为加

权权重，这里取 0.1，n 为句子的类别个数，b 为句子是否含有论文术语中的类别，取值为 0 或 1。句子分类可采用朴素贝叶斯、条件随机场等传统分类模型或基于 LSTM、GRU[27]等深度神经网络的分类器方案解决，这里不再赘述。

4　实验过程与结果分析

4.1　实验过程

为验证上述主题句识别方法的有效性，本文以气候变化领域数据进行实验。实验数据是以期刊为单位，从 Elsevier 上下载了包括 *Atmospheric Research* 等 10 种期刊论文全文数据共 31 430 篇。

从论文全文数据中提取论文的标题、摘要、大纲、全文等信息构建词向量训练语料，根据数据量及自然语言处理任务，选择合适的训练模型、优化算法及超参数。本文选择 Skip-gram 模型及 Hierarchical Softmax 方法，该搭配能更好地表示不常见的领域词汇，另外，其他超参数如上下文窗口为 5，词向量维度大小为 100 等。经过 5 个小时的训练，最终得到一个大小约为 400MB 的词向量文件。

对论文全文进行分句，提取该句子的位置、所在小节、大纲等文本特征，以每个小节为单位，利用上述词向量文件构建各句子间的基于 WMD 的语义相似度，采用改进 TextRank 算法识别主题句。测试预料采用由领域专家标注的 9 篇全文数据及其主题句。

4.2　结果分析

本文除了利用上述方法识别核心主题句，同时也基于相同的训练语料及测试文档，实现了传统 TextRank 算法、WMD 矩阵相加的方法、WMD 和 TextRank 算法，并对这 4 种方法进行分析比较，结果如表 1 所示。

表 1　气候变化领域 4 种算法的实验结果比较

方法	准确率	召回率	F1 值
TextRank	24.88%	22.94%	23.87%
WMD	22.89%	21.10%	21.96%
WMD+TextRank	23.38%	21.56%	22.43%
本文方法(WMD+TextRank +外部特征优化)	27.11%	25%	26.01%

表 1 的实验结果显示，虽然在同一文本段落中，本文方法较其他方法结果稍好一些，但这 4 种方法的结果均不太理想，通过对实验数据及结果进行分析发现，具体原因如下：

(1) 测试集均为与 El Niño 现象的数据，其标注内容与词向量训练语料相关度较低，部分词汇的词向量表达较差，影响了句子相关度计算；

(2) 测试集由一位领域专家标注，其评估准确度可能存在一定偏差；

(3) 本文以各小节为识别基本单元，采用固定比例方式进行识别，其假设论文的核心主题句在文中是均匀分布的。而事实上，论文核心主题句在论文各部分的分布是有差异的，例如引言、实验结果及结论部分体现着论文的主要产出成果，其核心主题句分布较高，而在对相关研究及方法介绍上，核心主题句的分布较低。采用同样比例的识别过程较大程度上影响了结果。

故本文使用计算机领域数据作为词向量训练语料，实验过程同上述过程一样，随机选择词向量训练语料中的 10 篇文章标记，标记数据上采用多人协同标注选取共同认可的句子，同时在识别过程中，加大论文首尾部分的识别比例为其他部分的两倍，最终实验结果如表 2 所示。

表 2　计算机领域 4 种算法的实验结果比较

方法	准确率	召回率	F1 值
TextRank	25.05%	38.59%	30.37%
WMD	20.24%	31.17%	24.54%
WMD+TextRank	27.66%	42.59%	33.54%
本文方法(WMD+TextRank +外部特征优化)	29.06%	44.75%	35.24%

通过对实验过程进行部分调整，其识别效果较上次实验 F1 值提高了近 10%，同时本文方法较传统的 TextRank 方法 F1 值提高了近 5%，取得了相对较好的效果。

对上述实验过程及结果进行总结，得出以下结论。

(1) 传统的 TextRank 算法识别效果比较稳定，其识别出的句子因受限于相似度计算方法，而普遍较长。

(2) 词向量的质量影响了句子相似度计算，从而影响了本文方法的识别结果。而对于传统的 TextRank 算法，其计算过程基于共现关系，而与词的潜在语义无关，所以词向量较差时效果稍好。

(3) 在句子可利用程度方面，虽然本文识别效果

仍有很大提升空间，但通过对本文方法所识别的结果进行认真分析发现，结果集中未命中的句子普遍也是有价值的句子，整体质量上较其他三种方法更好。

(4) 未命中的句子中，有一部分含有特殊引导词(如"Hence"、"In this paper"、"It shows that"等)、数值型文本等具有明确特征的句子，前者表明作者在论文中声明的重要总结性句子，而后者表明论文最具说服力的论据，两者在论证论文核心主题上起到重要作用。而由于采用图模型句子相似的方法，弱化了这些信息，使得这些句子未识别出来。可通过进一步研究来抽取这些特征，优化识别方法，提高识别效果。

综上所述，本文提出的主题句识别方法，不需要任何外部知识结构，利用全文构建词向量，改进相似度计算方法完善 TextRank 迭代过程，并基于外部特征对所得结果权值进行调整，同时利用文本段落句子内部联系以及论文整体外部结构的丰富信息，识别效果达到人工评测平均水平，较文献[12]要好，但仍有较大的改进之处。

5 结 语

本文分析了目前主流的主题句抽取方法，并针对科技论文的特点，基于领域词向量，融合 WMD 语义相似度的 TextRank 改进算法识别主题句。实验结果表明，本文的主题句识别方法能够较好地甄别科技论文小节内部中心句，辅以外部特征的权值调整后可以较好地识别出一篇论文的核心主题句，但在句子特征提取和词向量训练过程还存在一定不足，另外主题句识别方法中各项参数仍需进一步调优。下一步工作将继续优化方法中的各项参数，提取更有效的句子特征，并利用词向量发现核心词汇与句子的潜在关系，提升核心主题句识别准确率。

参考文献:

[1] Sunayama W, Yachida M. Panoramic View System for Extracting Key Sentences Based on Viewpoints and Application to a Search Engine[J]. Journal of Network and Computer Applications, 2005, 28(2): 115-127.

[2] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts [OL]. Unt Scholarly Works, 2004. https://digital.library.unt.edu/ark:/67531/metadc30962/m2/1/high_res_d/Mihalcea-2004-TextRank-Bringing_Order_into_Texts.pdf.

[3] Kusner M J, Sun Y, Kolkin N I, et al. From Word Embeddings to Document Distances[C]// Proceedings of the 32nd International Conference on Machine Learning. 2015: 957-966.

[4] Batcha N K, Aziz N A. An Algebraic Approach for Sentence Based Feature Extraction Applied for Automatic Text Summarization[J]. Journal of Computational & Theoretical Nanoscience, 2014, 20(1): 139-143.

[5] Luhn H P. The Automatic Creation of Literature Abstracts[J]. IBM Journal of Research and Development, 1958, 2(2): 159-165.

[6] Baxendale P B. Machine-Made Index for Technical Literature——An Experiment[J]. IBM Journal of Research and Development, 1958, 2(4): 354-361.

[7] 刘挺, 王开铸. 自动文摘的四种主要方法[J]. 情报学报, 1999, 18(1): 10-19. (Liu Ting, Wang Kaizhu. Four Kinds of Main Methods of Automatic Abstracting[J]. Journal of the China Society for Scientific and Technical Information, 1999, 18(1): 10-19.)

[8] Edmundson H P. New Methods in Automatic Extracting[J]. Journal of the ACM, 1969, 16(2): 264-285.

[9] Kupiec J, Pedersen J, Chen F. A Trainable Document Summarizer[C]//Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1995: 68-73.

[10] Conroy J M, O'leary D P. Text Summarization via Hidden Markov Models[C]//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2001: 406-407.

[11] Mihalcea R. Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization [C]//Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics, 2004: 20.

[12] 余珊珊, 苏锦钿, 李鹏飞. 基于改进的 TextRank 的自动摘要提取方法[J]. 计算机科学, 2016, 43(6): 240-247. (Yu Shanshan, Su Jindian, Li Pengfei. Improved TextRank-based Method for Automatic Summarization[J]. Computer Science, 2016, 43(6): 240-247.)

[13] 耿焕同, 蔡庆生, 赵鹏, 等. 一种基于词共现图的文档自动摘要研究[J]. 情报学报, 2005, 24(6): 651-656. (Geng Huantong, Cai Qingsheng, Zhao Peng, et al. A Kind of Automatic Text Keyphrase Extraction Method Based on Word Co-occurrence[J]. Journal of the China Society for Scientific and Technical Information, 2005, 24(6): 651-656.)

[14] 何维, 王宇. 基于句子关系图的网页文本主题句抽取[J]. 现代图书情报技术, 2009(3): 57-61. (He Wei, Wang Yu.

Extracting Topic Sentences from Web Text Based on Sentence Relationship Map[J]. New Technology of Library and Information Service, 2009(3): 57-61.)

[15] Saïd T, Evrard F. Intentional Structures of Documents[C]// Proceedings of the 12th ACM Conference on Hypertext and Hypermedia. ACM, 2001: 39-40.

[16] Harris Z S. Distributional Structure[A]. //Papers on Syntax[M]. Springer Netherlands, 1954.

[17] Bengio Y, Schwenk H, Senécal J S, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.

[18] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[19] Word2Vec [EB/OL]. [2016-12-26]. https://code.google.com/p/word2vec/.

[20] 郭庆琳, 李艳梅, 唐琦. 基于 VSM 的文本相似度计算的研究[J]. 计算机应用研究, 2008, 25(11): 3256-3258. (Guo Qinglin, Li Yanmei, Tang Qi. Similarity Computing of Documents Based on VSM [J]. Application Research of Computers, 2008, 25(11): 3256-3258.)

[21] Wang R, Neumann G. Recognizing Textual Entailment Using Sentence Similarity Based on Dependency Tree Skeletons[C]//Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007: 36-41.

[22] Wang D, Li T, Zhu S, et al. Multi-document Summarization via Sentence-level Semantic Analysis and Symmetric Matrix Factorization[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2008.

[23] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[OL]. arXiv Preprint. arXiv: 1301.3781, 2013.

[24] Ling H, Okada K. An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(5): 840-853.

[25] 何远标, 乐小虬, 张帆. 学术论文大纲中关键术语抽取方法研究[J]. 现代图书情报技术, 2014(3): 73-79. (He Yuanbiao, Le Xiaoqiu, Zhang Fan. Research on Keyphrase Extraction from Scholarly Article Outline[J]. New Technology of Library and Information Service, 2014(3): 73-79.)

[26] 何远标. 基于学术论文大纲的术语层级关系挖掘[D]. 北京: 中国科学院大学, 2014. (He Yuanbiao. Phrase Hierarchical Relationship Mining Based on Scholarly Article Outline[D]. Beijing: University of Chinese Academy of Sciences, 2014.)

[27] Chung J, Gulcehre C, Cho K H, et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling[OL]. arXiv Preprint. arXiv: 1412.3555, 2014.

作者贡献声明:

王子璇: 设计并实施技术方案、技术路线, 数据采集与清洗, 实验的分析和验证, 论文的起草、撰写以及最终版本修订;
乐小虬: 提出论文研究方向和主要研究思路, 优化研究方案及技术路线的设计, 论文修改;
何远标: 部分模块实现, 参与论文修改。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: lexq@mail.las.ac.cn。

[1] 王子璇, 乐小虬, 何远标. rec_result.xlsx. 气候变化领域和计算机领域主题句识别测试结果对比集.

收稿日期: 2017-01-19
收修改稿日期: 2017-03-13

Recognizing Core Topic Sentences with Improved TextRank Algorithm Based on WMD Semantic Similarity

Wang Zixuan[1, 2] Le Xiaoqiu[1] He Yuanbiao[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2] (University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] This paper aims to automatically recognize key sentences describing the research topics of scientific papers. [**Methods**] First, we used paper sections as the unit to organize sentence sets. Then, we calculated the WMD distance between sentences by trained domain word embeddings. Third, we optimized the iterative process of TextRank algorithm, and used external features to adjust sentence's weights. Finally, we identified the core topic sentences according to the sentence's weights descendingly. [**Results**] We examined the proposed method with scientific papers on climate changes and compared it with the traditional TextRank algorithm. The recognition efficiency (F-value) was about 5% higher than that of the TextRank algorithm. [**Limitations**] The extraction of sentence features needs to be improved, and word embedding training and related parameters of the proposed method need to be further optimized. [**Conclusions**] The improved TextRank algorithm, could effectively recognize inner core sentences of scientific paper sections. It could recognize core topic sentences of a paper with the adjusted weights of external features.

Keywords: WMD TextRank Semantic Similarity Topic Sentence Recognition External Features

ProQuest 和 CALIS 合作增加中文学术成果全球曝光度

ProQuest 和中国高校图书馆联盟——中国高等教育保障系统(CALIS)于近日宣布延长他们之间的长期合作关系。目前,来自中国知名大学的共 27 万篇学位论文摘要可在全球范围内进行获取, 这些记录以英文形式被 ProQuest 博硕士论文全球数据库(PQDT Global)收录, 世界各地 3 000 多所大学的科研工作者都能从中发现中国的学术研究成果。这一合作推动了全球科学研究, 帮助全球科研工作者更全面地了解和发现学术活动。同时也帮助了中国大学更好地向国外传播中国学生的科研工作。

这是 ProQuest 和 CALIS 就学位论文进行的第二个合作项目。此前, 他们已经共同合作了十几年。2003 年, 他们合作创建了一个论文资源库, 使 CALIS 成员图书馆能够在 CALIS 平台上发现和访问 PQDT Global 的 63 万多篇博硕士论文全文。

(编译自: http://www.proquest.com/about/news/2017/ProQuest-and-CALIS-Bring-Chinese-Scholarship-to-a-Global-Audience.html)

(本刊讯)